In Ordnung oder nicht in Ordnung, das ist hier die Frage!

Erklär-Qualität von Modellen

Das Modell kann aus dem Bereich maschinelles Lernen, künstliche Intelligenz oder anderen Datenanalyse-Methoden stammen. Bei der Erklär-Qualität geht es um die Chance, mit der ein Ergebnis (z. B. iO oder niO) korrekt durch das Modell vorhergesagt wird.

Konfusionsmatrix (confusion matrix)

Die Konfusionsmatrix ist eine Vierfeldertafel. Es wird gezählt, wie oft welche Kombination aus Realität (z. B. Teil iO oder Teil niO) und Modell-Vorhersage (z. B. Vorhersage iO oder Vorhersage niO) auftaucht.

Beispiel: automatische Schweißnaht-Prüfung

Sie erhalten ein Angebot für eine Anlage, in der mit Hilfe von künstlicher Intelligenz automatisch geprüft wird, ob eine Schweißnaht in Ordung (iO) oder nicht in Ordnung (niO) ist. Anhand von 254 Schweißnähten werden die Vorhersagen der KI-Anlage mit den Prüfergebnissen aus der bisherigen manuellen Prüfung verglichen (Tabelle 1).

Ist die Vorhersage gut genug, um die manuelle Prüfung durch die KI-Anlage zu ersetzen?[1]

Tabelle 1: Anzahl Prüfergebnisse KI-Anlage und manuelle Prüfung

Realität / **Modell**	Prüfergebnis: Teil ist iO	Prüfergebnis: Teil ist niO	gesamt
Vorhersage: Teil ist iO	**186**	**34**	220
Vorhersage: Teil ist niO	**12**	**22**	34
gesamt	198	56	254

[1] Die Antwort ist natürlich „Es kommt darauf an.“ Hier wird bewertet, wie schädlich falsche Entscheidungen sind bzw. welche Konsequenzen falsche Entscheidungen haben.

Allgemein werden die vier Kombinationsmöglichkeiten durch die englischen Abkürzungen TP, FP, TN und FN angegeben:

- TP: True Positive / Vorhersage ist iO & Prüfergebnis ist iO
- FP: False Positive / Vohersage ist iO & Prüfergebnis ist niO
- FN: False Negative / Vorhersage ist niO & Prüfergebnis ist iO
- TN: True Negative / Vorhersage ist niO & Prüfergebnis ist niO

Die Zuordnung „positiv“ und „negativ“ sind willkürlich. Typischerweise wird das Ergebnis, das identifiziert werden soll, als „positiv“ bezeichnet. Je nach Zuordnung kann das z. B. bei der KI-Anlage für das Schweißen „positiv = erkennt niO“ zuverlässig oder „positiv = erkennt iO zuverlässig“ bedeuten. Die Zuordnung muss deshalb bei der Auswahl geeigneter Kennzahlen berücksichtigt werden.

Tabelle 2: Konfusionsmatrix Zuordnung TP, FP, FN und TN

Realität / **Modell**	Prüfergebnis: Teil ist iO	Prüfergebnis: Teil ist niO
Vorhersage: Teil ist iO	**TP: True Positive**	**FP: False Positive**
Vorhersage: Teil ist niO	**FN: False Negative**	**TN: True Negative**

Kennzahlen für die Evaluation (evaluation metrics)

Die Kennzahlen für die Evaluation bei attributiven Zielgrößen wie beispielsweise iO/niO geben an, wie hoch die Erklär-Qualität eines Modells ist. Typischerweise werden hier englische Begriffe verwendet. Da sich sehr unterschiedliche deutsche Übersetzungen finden, sind die deutschen Begriffe lediglich in Klammern ergänzt.

Accuracy (Genauigkeit)

Die Accuracy gibt den Anteil der korrekten Vorhersagen gemessen an der Anzahl Vorhersagen insgesamt an.

$$\text{Accuracy} = \frac{\#TP + \#TN}{\#\text{Vorhersagen}} = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN}$$

#: Anzahl

Beispiel: Accuracy bei automatischer Schweißnaht-Prüfung

Accuracy: „Wie oft liefert das KI-Modell eine korrekte Vorhersage?“

Realität / Modell	Prüfergebnis: Teil ist iO	Prüfergebnis: Teil ist niO	gesamt
Vorhersage: Teil ist iO	**TP: 186**	**FP: 34**	220
Vorhersage: Teil ist niO	**FN: 12**	**TN: 22**	34
gesamt	198	56	254

$$\text{Accuracy} = \frac{\#TP + \#TN}{\#\text{Vorhersagen}} = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN} = \frac{186 + 22}{254} = 81{,}89\,\%$$

Precision (Präzision)

Die Precision gibt den Anteil der korrekten positiven Vorhersagen gemessen an der Anzahl positiven Vorhersagen insgesamt an.

$$\text{Precision} = \frac{\#TP}{\#TP + \#FP}$$

#: Anzahl

Beispiel: Precision bei automatischer Schweißnaht-Prüfung

Precision: „Wie oft ist die KI-Vorhersage „iO“ korrekt?“

Realität / Modell	Prüfergebnis: Teil ist iO	Prüfergebnis: Teil ist niO	gesamt
Vorhersage: Teil ist iO	**TP: 186**	**FP: 34**	220
Vorhersage: Teil ist niO	**FN: 12**	**TN: 22**	34
gesamt	198	56	254

$$\text{Precision} = \frac{\#TP}{\#TP + \#FP} = \frac{186}{186 + 34} = 84{,}55\,\%$$

Recall (Identifizierbarkeit positiver Ergebnisse, Empfindlichkeit)

Der Recall gibt den Anteil der korrekt identifizierten positiver Ergebnisse gemessen an der Anzahl positiver Vorhersagen durch das KI-Modell an.

$$\text{Recall} = \frac{\#TP}{\#TP + \#FN}$$

#: Anzahl

Beispiel: Recall bei automatischer Schweißnaht-Prüfung

Recall: „Wie oft wird ein iO-Teil korrekt durch die KI-Vorhersage als „iO“ eingestuft?“

Realität / **Modell**	Prüfergebnis: Teil ist iO	Prüfergebnis: Teil ist niO	gesamt
Vorhersage: Teil ist iO	**TP: 186**	**FP: 34**	220
Vorhersage: Teil ist niO	**FN: 12**	**TN: 22**	34
gesamt	198	56	254

$$\text{Recall} = \frac{\#TP}{\#TP + \#FN} = \frac{186}{186 + 12} = 93{,}94\,\%$$

Specificity (Identifizierbarkeit negativer Ergebnisse, Spezifität)

Die Specificity gibt den Anteil der korrekt identifizierten negativer Ergebnisse gemessen an der Anzahl negativer Vorhersagen durch das KI-Modell an.

$$\text{Specificity} = \frac{\#TN}{\#TN + \#FP}$$

#: Anzahl

Beispiel: Specificity bei automatischer Schweißnaht-Prüfung

Specificity: „Wie oft wird ein niO-Teil korrekt durch die KI-Vorhersage als „niO“ eingestuft?“

Realität / Modell	Prüfergebnis: Teil ist iO	Prüfergebnis: Teil ist niO	gesamt
Vorhersage: Teil ist iO	**TP: 186**	**FP: 34**	220
Vorhersage: Teil ist niO	**FN: 12**	**TN: 22**	34
gesamt	198	56	254

$$\text{Specificity} = \frac{\#TN}{\#TN + \#FP} = \frac{22}{22+34} = 39{,}29\,\%$$

F1-Score

Der F1-Score ergibt sich aus der Kombination von Recall und Precision. Er ist deshalb nicht mehr direkt inhaltlich interpretiertbar sondern eine allgemeine Kenngröße für die Erklär-Qualität eines Modells.

$$\text{F1-Score} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

#: Anzahl

Beispiel: F1-Score bei automatischer Schweißnaht-Prüfung

$$\text{F1-Score} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} = 2 \cdot \frac{93{,}94\,\% \cdot 84{,}55\,\%}{93{,}94\,\% + 84{,}55\,\%} = 89{,}00\,\%$$

Wann sind die Kennzahlen groß genug?

Es gibt keine allgemeinen Grenzwerte für die Höhe der Kennzahlen. Das hängt u. a. davon ab, wie teuer (Geld, verärgerte Kund:innen usw.) eine falsche Entscheidung ist, deshalb sollte neben den Kennzahlen immer bewertet werden, welche Konsequenzen aus einer falschen Entscheidung entstehen. Die Höhe der Kennzahlen alleine reicht nicht aus, um gute Entscheidungen für die praktische Anwendung zu treffen.